AFHRL-TP-89-41

AIR FORCE HUMAN RESOURCES LABORATORY

EVALUATION OF OPTIMAL APPROPRIATENESS MEASUREMENT FOR USE IN PRACTICAL SETTINGS

F. Drasgow
M. V. Levine
B. Williams
C. McCusker
G. L. Thomasson
R. G. Lim

Model Based Measurement Laboratory
University of Illinois
210 Education Building
1310 South Sixth Street
Champaign, Illinois 61820

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

January 1990
Final Technical Paper for Period October 1987- June 1989

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

JAMES A. EARLES
Contract Monitor

WILLIAM E. ALLEY, Technical Director
Manpower and Personnel Division

DANIEL L. LEIGHTON, Colonel, USAF
Chief, Manpower and Personnel Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	January 1990	Final - October 1987 to June 1989

4. TITLE AND SUBTITLE

Evaluation of Optimal Appropriateness Measurement for Use in Practical Settings

5. FUNDING NUMBERS

C - F41689-87-D-0012
PE - 62205F
PR - 2922
TA - 02
WU - 02

6. AUTHOR(S)

F. Drasgow
M. V. Levine
B. Williams
C McCusker
G. L. Thomasson
R. G. Lim

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Model Based Measurement Laboratory
University of Illinois
210 Education Building
1310 South Sixth Street
Champaign, Illinois 61820

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Manpower and Personnel Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5601

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

AFHRL-TP-89-41

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Optimal appropriateness indices were evaluated for use in applied settings. The first study demonstrated the adaptation of existing computer software to the problem of identifying cheaters with scores in specific test score ranges. It is shown that the Levine and Drasgow (1988) algorithm can be applied directly to this problem. Then simulated and real data were used to determine the rates of detection of cheating on 5, 10, or 15 items on each of two tests and obtaining a total test score in one of two total test score ranges. As expected, low rates of identification were obtained for cheating on only 5 items per test and reasonably high rates were obtained for cheating on 15 items. Moderately lower identification rates were obtained with real data than with simulation data. The extent of difficulties that are likely to occur when the assumptions of optimal indices are violated were evaluated in the four robustness experiments that constituted Study Two. Optimal indices were found to be quite robust to: misspecification of the ability density, the use of estimated item characteristic curves and option characteristic curves in place of the true (simulation) curves, and misspecification of the number of aberrant responses. On the other hand, substantially lower detection rates were obtained when local independence was violated by generating item responses with two correlated (.70) traits but computing optimal indices with the incorrect assumption of a unidimensional latent trait space. Some implications of these results are presented in the final section of this paper.

14. SUBJECT TERMS

Armed Services Vocational Aptitude Battery
appropriateness measurement
aptitude test

15. NUMBER OF PAGES
40

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

AFHRL-TP-89-41 January 1990

EVALUATION OF OPTIMAL APPROPRIATENESS MEASUREMENT FOR USE IN PRACTICAL SETTINGS

F. Drasgow
M. V. Levine
B. Williams
C. McCusker
G. L. Thomasson
R. G. Lim

Model Based Measurement Laboratory
University of Illinois
210 Education Building
1310 South Sixth Street
Champaign, Illinois 61820

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

Reviewed by

Linda T. Curran, Acting Chief
Enlisted Selection and Classification Function

Submitted for publication by

Lonnie D. Valentine, Jr., Chief
Force Acquisition Branch

SUMMARY

The military services have a vital concern in assuring that aptitude test scores used for enlistment selection and classification are appropriate measures of applicants' true abilities. Substantial bonuses have been paid to examinees with sufficiently high scores as enticements to enlist into selected occupations. Also, failures in the services' training schools due to a lower aptitude than that necessary for successful completion cost thousands of dollars per individual. Therefore, cheating to improve scores on an enlistment test is a threat to the integrity of the services' selection and training systems. The goal of appropriateness measurement is to identify individuals who have not been accurately assessed by a multiple-choice test and, therefore, preserve the integrity of the test.

This effort investigated the utility of several appropriateness indices in identifying cheaters who were very low or who were just below average in verbal and quantitative aptitudes. The amount of cheating was 5, 10, or 15 items on tests of approximately 50 items in length. Real data as well as data simulated to maximize realism were used in the investigation. Low rates of identification were obtained for cheating on 5 items. This was expected because on an item for which an examinee does not know the right answer, it is very difficult to distinguish a correct response due to cheating from a correct response due to a lucky guess. A small number of lucky guesses is not unusual. Reasonably high rates of identification were obtained when cheating occurred on 15 items.

The above findings were based on (a) the sample having a normal ability distribution, (b) known probabilities of correct responses, (c) cheaters having a fixed and known number of compromised items, and (d) a complete knowledge of which test items were verbal and which were quantitative. Some appropriateness indices worked reasonably well when actual examinee responding deviated from the first three conditions. Condition d cannot be violated; however, it is not necessary to develop a separate appropriateness measure for verbal and for quantitative aptitudes. A method for extending appropriateness measurement to two aptitude areas has already been developed and can be used when the items belonging to each aptitude area are designated.

It is concluded that the utilization of appropriateness indices for identification of examinees for retesting would be expected to improve the quality of a large testing program.

PREFACE

This effort was accomplished under Project 2922, Prototype Development and Validation of Selection and Classification Instruments. It represents the continuing effort of the Air Force Human Resources Laboratory to fulfill its research and development responsibilities through development and application of state-of-the-art methodologies in the area of enlisted selection and classification.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Standardized psychological tests are administered to tens of millions of examinees per year. One test, the Armed Services Vocational Aptitude Battery (ASVAB), is administered to approximately 2.5 million examinees annually. The scores that result from standardized tests affect the lives of examinees by opening and closing doors to training programs, employment, and education.

Appropriateness measurement was proposed by Levine and Rubin (1979) as a means for identifying individuals who have been mismeasured by a standardized test. A general approach to specifying statistically optimal methods for this task was recently presented by Levine and Drasgow (1988). Their approach can be used to determine appropriateness indices that are optimal in the sense that no other statistic computed from the same data can provide higher rates of detection of the specified testing anomaly at the same false positive rate.

Drasgow, Levine, and McLaughlin (1987, in press) and Drasgow, Levine, McLaughlin, and Earles (1987) compared optimal appropriateness indices to earlier, nonoptimal indices described by Drasgow, Levine, and Williams (1985), Rudner (1983), Sato (1975), Tatsuoka (1984), and Wright (1977). For unidimensional tests, they found that the best nonoptimal indices sometimes provided rates of detection of aberrant response patterns that were almost as high as the rates of optimal indices. In other cases, the best nonoptimal indices were far less powerful than optimal indices. Multi-test extensions of the nonoptimal indices were found to be less effective relative to multi-test optimal indices for a test battery consisting of two unidimensional tests. In this case, nonoptimal indices rarely provided rates of detection that were close to the detection rates of optimal indices.

A number of difficulties and uncertainties have limited applications of optimal appropriateness indices. To date, formulas for optimal indices have been derived for only a few types of mismeasurement. Some of the formulas that have been derived are quite complex. A considerable investment of time and effort has been necessary to develop and program algorithms for evaluating the complex formulas. Very little is known about the robustness of optimal indices to violations of their underlying assumptions.

The research reported here was conducted in response to these problems. In Study One, existing software was used to test specific hypotheses with optimal indices. The performance of optimal indices was evaluated and compared to nonoptimal indices. Study Two examined the robustness of optimal indices to four different violations of assumptions. Specifically, multidimensional item responses were analyzed with a unidimensional model, estimated item characteristic curves (ICCs) and option characteristic curves (OCCs) were used rather than the true ICCs and OCCs, ability parameters were sampled from a distribution related to the chi-square distribution with 10 degrees of freedom but optimal indices were computed assuming that ability

was normally distributed, and optimal indices were computed for forms of aberrance (e.g., cheating on 20% of the test) that did not match the way aberrance was simulated (e.g., cheating on 30% of the test).

Appropriateness Indices

The primary focus of the research described in this paper is the evaluation of optimal appropriateness measurement. In the next subsection, a brief summary of optimal indices is provided; references to articles containing technical details are also given. Results for two non-optimal appropriateness indices were also obtained in Study One. The first of these two indices is the standardized l_0 index, which was described by Drasgow et al. (1985). The second non-optimal index, F2, is a standardized fit statistic given by Rudner (1983).

Optimal appropriateness indices. Levine and Drasgow (1988) showed that a most powerful appropriateness index for a given form of aberrance on a unidimensional test is the likelihood ratio (LR) statistic

$$LR = \frac{\underline{P}_{Aberrant}(\mathbf{u})}{\underline{P}_{Normal}(\mathbf{u})} . \qquad (1)$$

Here $\underline{P}_{Aberrant}(\mathbf{u})$ denotes the likelihood of a vector of $\underline{n}$ item responses $\mathbf{u} = [\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_n]$ given a specified form of aberrance and $\underline{P}_{Normal}(\mathbf{u})$ denotes the likelihood of $\mathbf{u}$ given the model of normal responding.

To illustrate $\underline{P}_{Normal}(\mathbf{u})$ and $\underline{P}_{Aberrant}(\mathbf{u})$, assume that the item responses are scored dichotomously, the test is unidimensional, $\underline{P}_i(\theta)$ is the probability of a correct response to item $\underline{i}$ by normal examinees with ability θ, and the ability density is $\underline{f}(\theta)$. Then the conditional likelihood of $\mathbf{u}$ is

$$\underline{P}_{Normal}(\mathbf{u}|\theta) = \prod_{i=1}^{n} \underline{P}_i(\theta)^{u_i} [1 - \underline{P}_i(\theta)]^{1-u_i} \qquad (2)$$

and the marginal likelihood is

$$\underline{P}_{Normal}(\mathbf{u}) = \int \underline{P}_{Normal}(\mathbf{u}|\theta)\underline{f}(\theta)\underline{d}\theta . \qquad (3)$$

Levine and Drasgow (1988) showed that $\underline{P}_{Aberrant}(\mathbf{u})$ can also be computed as

$$\underline{P}_{Aberrant}(u) = \int \underline{P}_{Aberrant}(u|\theta)\underline{f}(\theta)d\theta \qquad (4)$$

and presented methods that allow $\underline{P}_{Aberrant}(u|\theta)$ to be computed fairly easily. A very efficient method for approximating the quantity in Equation 4 was devised by Levine (in preparation; see Drasgow, Levine, & McLaughlin, in press, for an application). Although Levine's approximation was developed in the context of a multidimensional test battery, it can also be used for unidimensional tests.

For a composite of two unidimensional tests, the likelihood is

$$\int\int P(U_1 = u_1|\theta_1)\, P(U_2 = u_2|\theta_2)\, f(\theta)d\theta, \qquad (5)$$

where $P(U_j = u_j|\theta_j)$ is the likelihood of the $\underline{n}_j$ item responses u_j on test $\underline{j}$, $\underline{j} = 1,2$, under either the normal or aberrant model. An interesting feature of Levine's approximation for either the unidimensional (Equation 4) or multidimensional (Equation 5) case is that the one- or two-dimensional integrals are evaluated without quadrature, thereby avoiding extremely intensive computations.

II. STUDY ONE
TESTING SPECIFIC HYPOTHESES

Purpose

Suppose a test administrator has the answer sheets from a set of examinees whose test scores just barely exceed a minimum threshold required to be hired, promoted, or admitted to a training program. Further, suppose it is known that some examinees earned their test scores honestly, while other examinees obtained the answers to some items prior to the exam and thus obtained passing scores by cheating. The task of the test administrator is to use each examinee's pattern of item responses to determine whether a passing score was obtained honestly.

Likelihood ratio

The test administrator should use the likelihood ratio given in Equation 1 to decide whether a passing score was obtained honestly because no other statistic computed from the item responses provides more accurate classification. To apply Equation 1 to the problem faced by the test administrator, $\underline{P}_{Normal}(u)$ would be interpreted as the likelihood of a response pattern u given that the examinee was responding honestly and $\underline{P}_{Aberrant}(u)$ would be interpreted as the likelihood of u given that the examinee was cheating. Stated simply, the likelihood ratio of Equation 1 compares the likelihood of u assuming that the examinee was cheating to the

likelihood of $\mathbf{u}$ assuming that the examinee was honest; a large likelihood ratio suggests that the examinee was in fact cheating.

For the test administrator to use Equation 1, there must be an explicit means for evaluating its numerator and denominator. In this subsection, it is shown how existing software can be used for this purpose.

A fact from elementary probability can be used to simplify the task of evaluating Equation 1. Specifically, suppose a set A is a subset of set B. Then

$$P(A|B) = P(A)/P(B). \tag{6}$$

Equation 6 can be derived from the usual formula for conditional probability $P(A|B) = P(A \text{ and } B)/P(B)$ because $P(A \text{ and } B) = P(A)$ when A is a subset of B.

Let ω denote the range of test scores that are subject to the test administrator's scrutiny. For example, ω might consist of the set of test scores that fall into the 50th to 54th percentiles. In addition, let $\underline{X}$ be the function that maps item responses into test scores. If number right scoring is used, for example,

$$\underline{X}(\mathbf{u}) = \underline{u}_1 + \underline{u}_2 + \ldots + \underline{u}_n .$$

Let $\mathbf{u}^*$ be a given sequence of responses such that $\underline{X}(\mathbf{u}^*)$ is in ω. With this notation, we can write the likelihood ratio that must be evaluated by the test administrator as

$$LR(\mathbf{u}^*) = \frac{\underline{P}_{\text{Aberrant}}(\mathbf{u}=\mathbf{u}^* | \underline{X}(\mathbf{u}) \text{ is in } \omega)}{\underline{P}_{\text{Normal}}(\mathbf{u}=\mathbf{u}^* | \underline{X}(\mathbf{u}) \text{ is in } \omega)} . \tag{7}$$

Applying Equation 6 to Equation 7 produces

$$LR(\mathbf{u}^*) = \frac{\underline{P}_{\text{Aberrant}}(\mathbf{u}=\mathbf{u}^*) \;/\; \underline{P}_{\text{Aberrant}}(\underline{X}(\mathbf{u}) \text{ is in } \omega)}{\underline{P}_{\text{Normal}}(\mathbf{u}=\mathbf{u}^*) \;/\; \underline{P}_{\text{Normal}}(\underline{X}(\mathbf{u}) \text{ is in } \omega)}$$

$$= \frac{\underline{P}_{\text{Aberrant}}(\mathbf{u}=\mathbf{u}^*)}{\underline{P}_{\text{Normal}}(\mathbf{u}=\mathbf{u}^*)} \cdot \underline{k} , \tag{8}$$

where $\underline{k}$ is a constant and thus can be ignored by the test administrator. Of course, this formula (and the specific $\underline{k}$) is valid only for patterns $\mathbf{u}^*$ with

$\underline{X}(\mathbf{u}^*)$ in ω. For such $\mathbf{u}^*$, $P_{\underline{X}|Normal}(\mathbf{u}=\mathbf{u}^*)$ can now be evaluated by Equations 2 and 3, and $P_{\underline{X}|Aberrant}(\mathbf{u})$ can be evaluated by Equation 4 and the methods described by Levine and Drasgow (1988) and Drasgow, Levine, and McLaughlin (in press).

Method

Overview. A study was conducted to examine the performances of optimal and non-optimal appropriateness indices on the task faced by the hypothetical test administrator. Both real and simulated data were analyzed in the study. The results obtained from the analysis of simulated data provide information about the performance of appropriateness indices under idealized conditions where all model assumptions are satisfied; the analysis of real data provides information about the indices' performances in operational conditions.

Data were generated to simulate normal responding to a test battery consisting of a test of verbal ability (V) and a test of quantitative ability (Q). In addition, data from presumably normal examinees responding to verbal and quantitative tests were analyzed. Response patterns with total test scores (V+Q) falling into two score ranges (20th through 24th percentiles and 50th through 54th percentiles) were selected. Compromise samples were formed by modifying either simulated response patterns or actual response patterns to simulate individuals who obtained total scores in the two score ranges by cheating. Appropriateness indices were computed for all response patterns, and rates of identification of the simulated cheaters were determined at various false positive rates.

The real data set, item characteristic curves, and option characteristic curves. The real data used in this study were from a sample of 13,571 examinees who responded to the ASVAB, version 17A, under operational conditions. To estimate item parameters, 3,392 examinees were chosen by selecting examinees 1, 5, 9, ... 13,569. A verbal test of 50 items was formed by combining the 35 item Word Knowledge test and the 15 item Paragraph Comprehension test. A quantitative test was formed by combining the 30 item Arithmetic Reasoning test and the 25 item Mathematics Knowledge test. The quantitative test contained 54 items after one Arithmetic Reasoning item was deleted because it was very easy (its item difficulty parameter was not accurately estimated).

Three-parameter logistic item characteristic curves were estimated by the method of marginal maximum likelihood with the BILOG (Mislevy & Bock, 1984) computer program. Non-parametric estimates of ICCs and option characteristic curves based on Levine's (1985, 1989a, 1989b) Multilinear Formula Score (MFS) theory were obtained using the ForScore computer program (Williams & Levine, in preparation). Additional details about the non-parametric estimates were given by Lim, Williams, McCusker, Mead, Thomasson, Drasgow, and Levine (1989). The estimated three-parameter logistic ICCs and the estimated non-parametric ICCs and OCCs were used in all subsequent analyses of the real data.

To maximize the realism of the simulation portion of this study, ICCs and OCCs that had been estimated from the ASVAB data set were used as the "true" (i.e., simulation) ICCs and OCCs rather than an arbitrarily specified

set of item parameters. This choice of ICCs and OCCs increases the comparability of the results obtained from the simulation and real data.

Item response models. In the portion of Study One that analyzed the actual ASVAB data, examinees' item responses were scored either dichotomously or polychotomously, and appropriateness indices were computed with either the three-parameter logistic ICCs or multilinear formula scoring ICCs and OCCs. Specifically, appropriateness indices were computed with the following item scoring and item response models:

1. dichotomously scored responses analyzed with three-parameter logistic ICCs;

2. dichotomously scored responses analyzed with multilinear formula scoring ICCs;

3. polychotomously scored responses analyzed with multilinear formula scoring ICCs and OCCs.

For the simulation portion of Study One, data were generated for each of the three conditions listed above (e.g., three-parameter logistic ICCs were used to generate dichotomous item responses). Appropriateness indices were then computed with the model used to generate each sample, which yielded analyses of simulated data that were parallel to the analyses of real data.

Percentiles. The following procedure was used to determine the total test scores corresponding to the 20th, 24th, 50th, and 54th percentiles for Study One. First, the estimated three-parameter logistic ICCs were used to generate 100,000 response patterns by the process for simulating normal response patterns (see below). Next, number-right scores were computed for each simulated verbal and quantitative test. Number-right scores on these two tests were then separately standardized and a total score was computed as the sum of the two standardized scores. Finally, the frequency distribution of the total score was tabulated and used to determine the values of the total test score association with specific percentiles.

Simulated normal response patterns. For each of the three item response models listed above, a simulated normal response pattern (i.e., a non-cheater) was created by sampling $\boldsymbol{\theta} = [\theta_1, \theta_2]$ from the standardized bivariate normal distribution with correlation .7. θ_1 was used with the simulation ICCs and OCCs for the verbal test to generate locally independent item responses. Similarly, θ_2 was used to generate locally independent item responses for the quantitative test. Response patterns were repeatedly generated until 4,000 simulated examinees were collected for the low score range (20th through 24th percentiles) normal sample and for the moderate score range (50th through 54th percentiles) normal sample.

Real normal response patterns. Real normal response patterns were obtained by first selecting each response pattern that was not included in the sample used to estimate ICCs and OCCs (i.e., response patterns were taken from the magnetic tape containing 13,571 response patterns, but the 3,392 patterns used for item calibration were excluded). Next, a total test score was computed for each response pattern in the manner described previously. Response patterns with total test scores in either the low

score range or the moderate score range were then written to separate files. A total of 480 response patterns had total test scores in the 20th through 24th percentiles and 533 response patterns had total test scores in the 50th through 54th percentiles.

Spuriously high manipulation applied to simulated data. Cheating was simulated by first generating a normal response pattern and then rescoring k item responses to be correct, regardless of the original response. The rescored items were randomly selected for each response pattern, and so Levine and Drasgow's (1988) method for evaluating $P_{Aberrant}(u)$ could be applied directly.

Response patterns were generated with 5, 10, or 15 items per test rescored to simulate cheating. This process was continued until 2,000 response patterns with total scores in the low score range and moderate score range were collected. An attempt was made to generate 18 samples by factorially crossing the three item response models, the three levels of simulated cheating (5, 10, or 15 items per test), and the two score ranges (20th through 24th percentiles and 50th through 54th percentiles); however, the 15 item spuriously high manipulation consistently produced response patterns with total scores that exceeded the 24th percentile. Consequently, it was possible to obtain only 15 spuriously high samples.

Spuriously high manipulation applied to real data. Only response patterns not used for item calibration and not in either normal sample were subjected to the spuriously high manipulations. The 5, 10, and 15 item spuriously high manipulations were applied to each of these response patterns, and a response pattern was selected if its total score fell in either the low or moderate score ranges. A total of 524, 635, and 654 response patterns were obtained for the moderate score range in the 5, 10, and 15 item spuriously high conditions. For the low score range, 408 and 310 response patterns were obtained in the 5 and 10 item conditions. Again, the 15 item spuriously high manipulation produced response patterns with test scores above the 24th percentile.

Analysis. Optimal appropriateness indices were computed for the samples of simulated and real normal response patterns using the Levine and Drasgow (1988) algorithm for spuriously high responding to 5, 10, and 15 items per test. Correctly specified optimal indices were always computed; for example, the optimal index for 10 spuriously high responses per test was computed for aberrant response patterns that had been subjected to this manipulation. The non-optimal indices were also computed for each normal and aberrant sample.

After computing appropriateness indices, receiver operating characteristic (ROC) curves were constructed. These curves depict the proportions of the response patterns in an aberrant sample that can be identified at various false positive rates. Of course, it is desirable to have a high detection rate (i.e., a high proportion of aberrant response patterns detected) at a low false positive rate.

Results

Rates of detection of simulated cheating for the low score range are presented in Table 1 for the simulated data. From Table 1 it is evident that simulated cheating on five items per test was very difficult to detect: Only 26% of the simulated cheaters were detected by the most sophisticated analysis when the false positive rate was 5%. The optimal index computed for the three-parameter logistic model was able to identify just 25%. Table 1 shows that cheating on 10 items per test was much easier to identify; for example, the optimal index for the MFS analysis of polychotomously scored responses identified 67% of the simulated cheaters at a false positive rate of 5%. The detection rates were 61% and 60% when the responses were scored dichotomously and analyzed with MFS and three-parameter logistic optimal methods.

Table 1 shows that the non-optimal ℓ_o and F2 indices had detection rates modestly below the detection rates of optimal indices for dichotomously scored responses. Their rates of detection rather substantially trailed the rates provided by the MFS optimal index for polychotomous scoring.

Table 2 presents results for actual ASVAB response patterns that had been modified to simulate individuals who obtained scores in the 20th through 24th percentile by cheating. Comparing the results for simulation data summarized in Table 1 to the real data results in Table 2 shows generally lower detection rates for real data. A word of caution is needed here: It was not possible to use samples of the size that ensure inconsequential sampling fluctuations (say, 4,000 normals and 2,000 aberrants) from the ASVAB data set. Thus, the numbers contained in Table 2 are subject to rather large sampling errors. Candell and Levine (1989) provide details about the expected sizes of sampling errors of ROC curves).

Two explanations for the lower detection rates in Table 2 are readily available. First, model misspecifications of various kinds may have had detrimental effects. This explanation was examined in Study Two, which was conducted to evaluate the consequences of a variety of misspecifications. A second explanation of the lower detection rates in Table 2 is that the normal sample used to determine false positive rates was not entirely normal. This sample, which consisted of actual ASVAB response patterns, might have contained a few truly aberrant response patterns. As one check of this latter hypothesis, the magnitudes of the likelihood ratios for 5% false positive rates where determined for the normal samples used in the simulation analyses and in the ASVAB analyses. Optimal indices were computed given the (incorrect) assumption that there were 10 spuriously high responses per test. The likelihood ratios are:

	Poly. MFS	Dichot. MFS	3PL
Simulation normal sample	2.10	2.21	2.19
ASVAB normal sample	5.36	4.24	3.07

Table 1. Selected Rates of Detection of Spuriously High Response Patterns with Total Test Scores in the 20th Through 24th Percentile, Simulation Data

False Pos. Rate	Test	Polychot. MFS Optimal	Dichot. MFS Optimal	Dichot. MFS ℓ_o	Dichot. MFS F2	3PL Optimal	3PL ℓ_o	3PL F2
			5 Spuriously High Responses Per Test					
.001	V	00	01	00	00	00	00	00
	Q	01	01	00	01	02	00	01
	MT	01	01	01	01	03	02	02
.01	V	05	04	02	02	05	03	02
	Q	07	05	04	04	06	05	05
	MT	10	08	05	06	08	06	07
.03	V	11	10	06	05	10	06	05
	Q	14	12	10	11	13	11	11
	MT	20	17	12	12	18	15	14
.05	V	15	14	09	09	13	10	09
	Q	22	19	16	16	18	17	17
	MT	26	24	18	19	25	20	20
.10	V	24	22	19	18	23	21	20
	Q	34	30	27	28	30	29	29
	MT	42	38	30	31	36	30	30
			10 Spuriously High Responses Per Test					
.001	V	05	04	03	01	03	02	02
	Q	04	05	04	04	08	03	04
	MT	10	15	10	10	19	11	09
.01	V	17	13	10	06	15	09	07
	Q	25	19	17	15	19	18	15
	MT	44	34	25	26	37	26	26
.03	V	28	23	19	14	24	20	17
	Q	42	35	32	32	33	31	29
	MT	59	51	40	39	52	44	43
.05	V	35	29	24	21	31	27	23
	Q	52	45	40	39	43	39	38
	MT	67	61	50	49	60	51	50
.10	V	46	41	38	34	42	39	38
	Q	67	61	56	55	60	55	55
	MT	79	74	64	64	73	64	62

Table 2. Selected Rates of Detection of Spuriously High Response Patterns with Total Test Scores in the 20th Through 24th Percentile, Real Data

False Pos. Rate	Test	Polychot. MFS	Dichot. MFS			3PL		
		Optimal	Optimal	ℓ_0	F2	Optimal	ℓ_0	F2
		5 Spuriously High Responses Per Test						
.001	V	00	00	00	00	00	01	00
	Q	02	00	00	01	01	00	00
	MT	01	00	00	00	02	00	00
.01	V	04	03	01	00	02	01	01
	Q	03	04	02	03	03	04	05
	MT	05	06	02	02	07	01	02
.03	V	10	10	04	02	09	04	02
	Q	07	05	05	06	08	08	07
	MT	14	13	06	07	14	10	07
.05	V	14	15	08	04	14	07	03
	Q	15	10	11	08	13	11	10
	MT	19	16	15	14	20	17	17
.10	V	26	21	12	12	21	13	11
	Q	24	21	21	19	25	24	21
	MT	32	33	24	22	31	26	21
		10 Spuriously High Responses Per Test						
.001	V	02	02	01	00	01	01	00
	Q	08	07	01	06	07	02	05
	MT	14	00	00	00	05	00	00
.01	V	08	05	02	00	05	03	01
	Q	17	14	15	12	16	20	20
	MT	15	16	09	07	22	09	07
.03	V	16	15	09	05	16	10	05
	Q	32	27	25	21	32	31	25
	MT	45	40	22	24	47	29	26
.05	V	28	23	15	07	19	14	07
	Q	42	43	34	28	43	37	34
	MT	52	53	36	36	58	45	43
.10	V	40	37	25	21	32	27	17
	Q	52	53	48	44	55	51	47
	MT	68	68	55	53	66	59	52

The likelihood ratio is the ratio of the likelihood of a response pattern given the model for aberrant responding--10 spuriously high responses per test--to the likelihood of the response pattern given the model for normal responding. A large likelihood ratio indicates that the model for aberrant responding "explains" the response pattern better than the normal model. The likelihood ratios shown above imply that the model for aberrant responding provides a good fit (relative to the model for normal responding) for more nominally normal ASVAB response patterns than simulation normal (and hence truly normal) response patterns. Note further that the optimal index is targeted for a specific form of aberrance (spuriously high responding), unlike goodness of fit indices such as ℓ_0 and F2 that test for any departure from normal responding. Thus, these results are consistent with the hypothesis that some ASVAB examinees may have received coaching.

Detection rates for simulated data with total test scores in the moderate score range are shown in Table 3. Again it was very difficult to identify response patterns that had been subjected to the five items per test spuriously high manipulation. One reason for this difficulty is that the version of the Levine and Drasgow (1988) algorithm used in this study makes no assumptions about <u>which</u> items were compromised; all items were assumed to be equally likely candidates for cheating. It seems likely that higher detection rates would be obtained if more were known about the relative likelihood of cheating on each item. For example, if new items introduced in a test administration or otherwise known to be secure can be assumed to have zero probability of spurious responses, then detection rates can be significantly increased by utilizing a more general version of the Levine and Drasgow algorithm. For another example, if the response options for some items are reordered because it is suspected that some examinees have memorized the answer key, the more general Levine and Drasgow (1988) algorithm can incorporate this additional information.

Table 3 shows moderate detection rates for cheating on 10 items per test and high detection rates for cheating on 15 items per test. Specifically, the best index identified 70% of the cheaters in this latter condition when the false positive rate was 5%. The detection rates for the two optimal indices computed with dichotomously scored responses were 62% and 62%. The non-optimal indices detected roughly 40% at a 5% false positive rate; the optimal index for polychotomous scoring achieved a somewhat higher detection rate at a false positive rate of only 1%.

A generally similar pattern of results was obtained in the analysis of the actual ASVAB data. Table 4 shows that it is a difficult task to identify cheating by near average ability examinees on a small to moderate number of items (5 or 10 items). Even the best appropriateness indices detect no more than 30% of such response patterns at a 5% false positive rate. These aberrant response patterns are difficult to identify because a substantial number of items were answered correctly <u>before</u> the spuriously high manipulation was applied. Thus, the aberrance manipulation does <u>not</u> produce a particularly unusual response pattern, namely one with several correct answers to hard items juxtaposed with incorrect answers to easy items.

Table 3. Selected Rates of Detection of Spuriously High Response Patterns with Total Test Scores in the 50th Through 54th Percentile, Simulation Data

False Pos. Rate	Test	Polychot. MFS Optimal	Dichot. MFS Optimal	Dichot. MFS ℓ_o	Dichot. MFS F2	3PL Optimal	3PL ℓ_o	3PL F2
		5 Spuriously High Responses Per Test						
.001	V	00	00	00	00	00	00	00
	Q	01	01	00	01	01	01	01
	MT	01	00	00	00	00	00	00
.01	V	03	03	02	01	03	02	01
	Q	04	03	03	03	03	03	03
	MT	05	04	02	02	05	04	03
.03	V	07	07	04	03	08	05	03
	Q	09	08	08	06	09	09	00
	MT	12	09	07	07	10	09	09
.05	V	10	10	07	05	12	08	05
	Q	13	13	10	11	14	12	12
	MT	17	15	11	11	16	12	12
.10	V	19	19	14	12	21	15	13
	Q	23	22	21	20	23	20	21
	MT	28	26	20	19	27	21	21
		10 Spuriously High Responses Per Test						
.001	V	01	01	00	00	01	01	00
	Q	03	03	01	01	02	02	01
	MT	05	01	02	01	02	01	01
.01	V	07	06	02	01	05	03	01
	Q	14	08	07	06	10	10	08
	MT	18	12	06	07	14	08	07
.03	V	13	12	06	04	13	07	04
	Q	26	20	16	15	21	20	19
	MT	32	27	14	15	26	18	18
.05	V	18	17	10	07	18	10	06
	Q	32	28	21	21	29	26	25
	MT	41	36	21	23	35	24	25
.10	V	28	28	17	14	27	19	15
	Q	46	42	36	36	44	37	37
	MT	56	52	33	34	49	36	38

Table 3 (concluded)

False Pos. Rate	Test	Polychot. MFS Optimal	Dichot. MFS Optimal	Dichot. MFS ℓ_0	Dichot. MFS F2	3PL Optimal	3PL ℓ_0	3PL F2
			15 Spuriously High Responses Per Test					
.001	V	07	05	01	00	05	03	00
	Q	09	06	03	04	05	06	05
	MT	21	08	06	03	07	05	05
.01	V	16	14	05	01	15	08	01
	Q	29	22	17	16	25	20	18
	MT	46	33	17	16	36	23	19
.03	V	28	25	11	06	27	14	06
	Q	48	39	31	30	41	34	36
	MT	62	53	31	30	53	36	36
.05	V	34	30	17	10	34	20	11
	Q	56	47	37	38	51	42	42
	MT	70	62	40	41	62	44	44
.10	V	46	42	27	20	44	32	23
	Q	69	61	52	53	65	53	54
	MT	82	74	54	54	75	57	59

Table 4. Selected Rates of Detection of Spuriously High Response Patterns with Total Test Scores in the 50th Through 54th Percentile, Real Data

False Pos. Rate	Test	Polychot. MFS	Dichot. MFS			3PL		
		Optimal	Optimal	ℓ_o	F2	Optimal	ℓ_o	F2
		5 Spuriously High Responses Per Test						
.001	V	00	00	00	00	00	00	00
	Q	00	00	00	00	02	00	00
	MT	00	00	01	00	01	00	00
.01	V	00	01	01	01	02	02	01
	Q	03	01	02	01	03	02	02
	MT	01	02	02	01	04	04	03
.03	V	02	06	04	02	05	05	04
	Q	08	09	07	06	09	06	06
	MT	07	09	06	06	13	07	07
.05	V	07	10	07	06	08	08	06
	Q	11	11	11	10	14	12	10
	MT	11	12	11	09	17	12	10
.10	V	16	15	11	10	15	13	09
	Q	21	21	20	19	22	20	21
	MT	20	24	20	18	23	20	20
		10 Spuriously High Responses Per Test						
.001	V	00	00	00	00	00	00	00
	Q	02	02	00	00	06	01	00
	MT	02	02	00	01	03	01	01
.01	V	01	02	01	00	03	01	00
	Q	06	08	05	03	09	05	04
	MT	07	09	06	04	12	08	05
.03	V	06	08	04	03	10	07	05
	Q	18	18	14	14	19	13	15
	MT	16	22	12	13	22	14	15
.05	V	12	15	08	06	13	10	07
	Q	23	21	20	18	24	21	19
	MT	26	28	19	18	30	20	19
.10	V	22	23	14	10	23	16	11
	Q	36	32	31	29	34	31	31
	MT	41	37	29	29	40	33	32

Table 4 (concluded)

False Pos. Rate	Test	Polychot. MFS Optimal	Dichot. MFS Optimal	Dichot. MFS ℓ_0	Dichot. MFS F2	3PL Optimal	3PL ℓ_0	3PL F2
		15 Spuriously High Responses Per Test						
.001	V	02	02	02	00	01	03	00
	Q	04	04	03	00	10	05	01
	MT	08	04	03	02	09	04	04
.01	V	08	08	04	00	08	05	01
	Q	20	16	12	07	18	12	10
	MT	27	24	14	08	25	18	12
.03	V	16	22	09	04	23	12	08
	Q	31	31	24	23	30	24	25
	MT	41	43	25	25	44	28	29
.05	V	27	27	15	09	29	18	10
	Q	41	35	33	30	40	34	30
	MT	50	53	36	33	54	39	36
.10	V	37	39	24	17	39	26	18
	Q	57	51	44	43	53	44	46
	MT	65	66	51	50	68	53	52

The rates of detection of response patterns subjected to the 15 item per test spuriously high manipulation are moderately high. For example, about 50% of these patterns are detected at a 5% false alarm rate. This higher detection rate is of course in part due to the severity of the manipulation. But, an important additional ingredient is that prior to the spuriously high manipulation the response patterns were indicative of fairly low ability. Thus, the patterns contained some incorrect answers to easy items. When the spuriously high manipulation resulted in correct answers to some of the harder items, detection of the simulated cheating was possible.

Rates of detection are somewhat lower in Table 4 than in Table 3, which again may be due to one of the forms of model misspecification examined in Study Two or due to the inclusion of truly aberrant response patterns in the nominally normal ASVAB sample. Likelihood ratios yielding a 5% false positive rate were determined for the ASVAB and simulation normal samples given the assumption of 10 spuriously high responses per test. The likelihood ratios are:

	Poly. MFS	Dichot. MFS	3PL
Simulation normal sample	4.10	3.94	3.86
ASVAB normal sample	7.60	5.73	5.17

As with the lower ability range, the likelihood ratios suggest that some aberrant response patterns may have been included in the nominally normal ASVAB sample.

III. STUDY TWO
ROBUSTNESS OF OPTIMAL INDICES TO VIOLATIONS OF ASSUMPTIONS

Purpose

There are a variety of violations of the optimal indices' assumptions that could create problems in operational settings. These violations include:

1. the use of estimated ICCs and OCCs in place of the true ICCs and OCCs;

2. violations of local independence that surely occur in real data;

3. differences between the assumed ability density in Equation 5 and the true ability density.

In addition to these three forms of model misspecification, another kind of misspecification is sure to occur in operational settings. The Levine and Drasgow (1988) algorithm assumes that the number of spuriously high or spuriously low responses on each test is known. However, such information is not usually available when a test is administered to examinees who may have been coached in a variety of ways. Thus, a fourth model

misspecification consists of violations of the assumed number of spurious responses per test.

Each of these four model misspecifications was investigated in Study Two. In each case, a misspecified index was computed in addition to the truly optimal index. Comparing the detection rates of the truly optimal index to the misspecified index shows the impact of the misspecification.

Method

Item characteristic curves and option characteristic curves. Although Study Two was entirely a simulation study, it was desirable to make the simulation as realistic as possible. For this reason, the very accurate estimates of item and option characteristic curves were obtained for the ASVAB items from Study One.

To this end, response patterns 1, 3, 5,... were initially selected from the complete sample, yielding a total of 6,785 patterns. To reduce this sample to a more manageable size, but still obtain very accurate ICC and OCC estimates, some examinees with average abilities were excluded whereas all examinees with extreme abilities were retained. (Estimation of ICCs and OCCs is typically very accurate for moderate ability ranges, but far less accurate in extreme ability ranges.) To avoid systematically violating local independence, response patterns were excluded on the basis of their scores on the 35 item General Science (GS) test rather than the verbal or quantitative tests. Response patterns with GS number-right scores of 15, 17, or 19 were deleted. This left a sample of 5,301 patterns, as 503, 518, and 463 patterns had scores of 15, 17, and 19, respectively.

As in Study One, marginal maximum likelihood estimates of the item parameters of the three-parameter logistic model were obtained with the BILOG (Mislevy & Bock, 1984) computer program and non-parametric estimates of ICCs and OCCs based on Levine's (1985, 1989a, 1989b) MFS theory were obtained with the ForScore computer program. Fit plots showed very accurate modeling of empirical proportions for the multilinear formula scoring ICCs and OCCs. Figure 1 shows a typical fit plot; the multilinear formula scoring estimate of the ICC is given by the dashed line in the upper left panel; the solid lines in the other three panels show conditional OCCs (OCCs divided by $[1-\underline{P}_i(\bullet)]$) for the three incorrect options.

Samples and analyses. The following general process was used to evaluate the effects of each of the four forms of misspecification described above. First, a normal sample of 4,000 response patterns was generated with the ICCs and OCCs described above. Then (except for the misspecified aberrance condition) two samples of 2,000 aberrant response patterns were generated, again with the ICCs and OCCs estimated from the sample of 5,301. One sample contained normal response patterns that had been subjected to the 10 item per test spuriously high manipulation, and the other sample contained patterns subjected to the 10 item per test spuriously low manipulation. Four aberrant samples of 2,000 patterns were created for the aberrance misspecification condition. Here samples were created with 5 and 15 item per test spuriously high manipulations and with 5 and 15 item per test spuriously low manipulations.

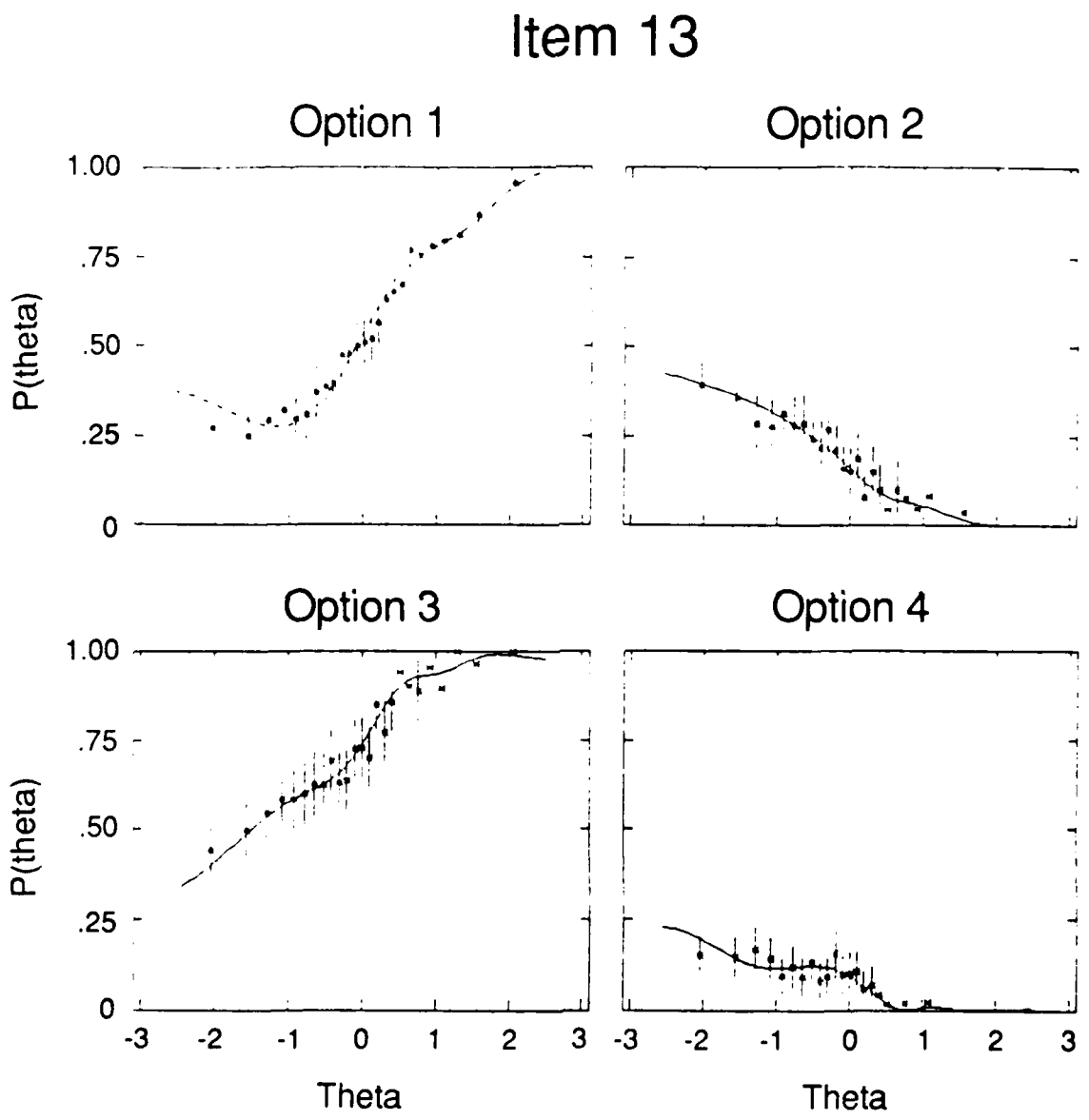

Figure 1. Fit Plots for an Item Characteristic Curve and Three Conditional Option Characteristic Curves Obtained with the ForScore Computer Program.

For three of the misspecifications, $\boldsymbol{\theta} = |\theta_1, \theta_2|$ was sampled from the standardized bivariate normal distribution with correlation .7. The sampling of $\boldsymbol{\theta}$ values in the misspecified ability density condition is described below. Note that there was no selection of response patterns as in Study One; all normal and aberrant response patterns were included.

A separate analysis was conducted to evaluate each of the four forms of misspecification. In each case, correctly specified optimal indices were computed as well as incorrectly specified optimal indices.

The first form of misspecification consisted of computing optimal indices with estimated ICCs and OCCs in place of the true ICCs and OCCs. To examine the effects of this substitution, the multilinear formula scoring ICCs and OCCs were used to simulate a test calibration sample of 3,000 response patterns. Then multilinear formula scoring ICCs and OCCs were estimated from this sample of 3,000 using the ForScore program and three-parameter logistic ICCs were estimated with the BILOG program. Finally, optimal appropriateness indices were computed for the normal and aberrant response patterns described above using the correct ICCs and OCCs as well as the estimated (from the simulated calibration sample of 3,000) ICCs and OCCs.

Note that the multilinear formula scoring ICCs and OCCs estimated from the simulation sample of 3,000 response patterns differ from the simulation ICCs and OCCs only to the extent of estimation error. In contrast, the three-parameter logistic ICCs estimated from the sample of 3,000 differ from the simulation ICCs both because of estimation errors and the fact that the true ICCs were not exactly three-parameter logistic. It seemed reasonable to incorporate this latter type of misspecification for the three-parameter logistic because ICCs are not necessarily correctly modelled by curves in the three-parameter logistic family.

The second form of misspecification investigated in Study Two consisted of violations of local independence. As described previously, item responses were generated to simulate a two-dimensional test where the two latent traits had a correlation of .7. The misspecified optimal indices made the incorrect assumption that the entire item pool of 104 items was unidimensional. Then optimal indices for a single long unidimensional test were computed in the misspecification condition; the correctly specified multi-test optimal indices were also computed

A misspecified ability density was the third form of misspecification studied. In earlier research (e.g., Drasgow, Levine, McLaughlin, & Earles, 1987), the ability density $\underline{f}(\bullet)$ in Equation 5 has been taken as the standard normal. This density is undoubtedly incorrect for a population of examinees when there has been self-selection or some other selection prior to administration of the exam (e.g., when recruiters prescreen applicants).

To simulate ability density misspecification, two numbers $\underline{X}$ and $\underline{Y}$ were sampled from a truncated chi-square distribution with 10 degrees of freedom (the bottom .01% and top 1.4% of the distribution were discarded since multilinear formula scoring ICCs and OCCs were defined only for θs less than 3 in absolute value). Then θ_1 was taken as $[\underline{X} - E(\underline{X})]/\sqrt{Var(\underline{X})}$ (i.e., a standardized version of the truncated chi-square). The density of θ_1 is

shown in Figure 2, along with the standard normal density. θ_2 was constructed by first standardizing $\underline{Y}$ and then computing $\theta_2 = \underline{a}\theta_1 + (1-\underline{a})\underline{z}_y$, where $\underline{z}_y$ is the standardized $\underline{Y}$ and $\underline{a} = .4995$ was chosen so that θ_1 and θ_2 had a correlation of .7. Finally, misspecified optimal indices were computed with the incorrect assumption that $[\theta_1,\theta_2]$ was sampled from the standardized bivariate normal distribution with correlation .7. Correctly specified optimal indices were also computed.

The final misspecification concerned the number of aberrant responses made by an examinee. Test administrators ordinarily do not know how many item responses might be aberrant. To evaluate the performances of optimal indices under these conditions, response patterns with 5 or 15 aberrant responses per test were created, and then the optimal index for 10 aberrant responses per test was computed as well as the correctly specified optimal index.

Results

True versus estimated ICCs and OCCs. Table 5 presents selected detection rates of spuriously high and spuriously low response patterns for the ICC and OCC misspecification condition. From this table it is evident that only minimal reduction in detection rates occurred as a result estimation error. The greatest shrinkage was expected for the polychotomous MFS analysis; here the detection rates for optimal indices computed for true and estimated ICCs and OCCs were 85% and 82% in the spuriously low condition and 39% and 36% in the spuriously high condition when the false positive rate was 5%. This small amount of shrinkage clearly indicates that the effects of the estimation errors obtained with a calibration sample of 3,000 were generally inconsequential.

There is one discrepant value in Table 5: When the false positive rate was .001, the detection rate for the polychotomous MFS multi-test optimal index was much lower for estimated ICCs and OCCs in the spuriously low condition. Although this result may be due to errors of estimation of the ICCs and OCCs, it may also be due to the fact that Table 5 presents empirical detection rates (i.e., the numbers in Table 5 would be different if we replicated our analysis but used a different seed for the random number generator). The cutting score for classification is determined from only 4 normal response patterns when the false positive rate is .001; this cutting score is likely to have considerable sampling error.

Very little decrement in detection rates is evident in the dichotomous MFS analysis. This finding corroborates results obtained by Levine, Drasgow, Williams, McCusker, and Thomasson (under review), who found very small estimation errors with their "ideal observer" methodology (i.e., an observer who uses an optimal statistical procedure to distinguish response patterns generated from true versus estimated ICCs).

Finally, the detection rates for the estimated three-parameter logistic ICCs are nearly as high as the rates for the dichotomous analysis with the true multilinear formula scoring ICCs. From this finding it appears that the joint effects of estimation errors and departures from the three-parameter logistic parameter form were generally inconsequential. Note,

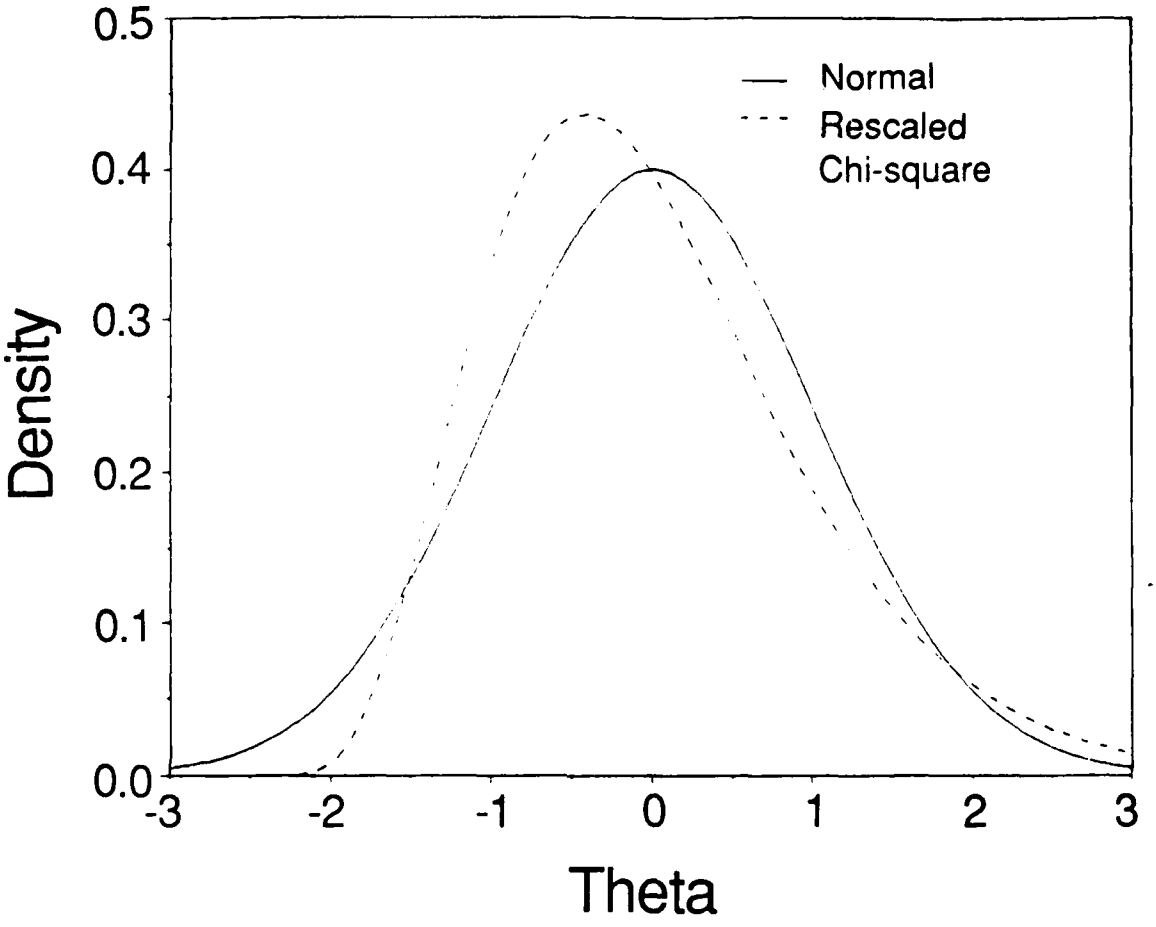

Figure 2. Density Functions of the Rescaled Chi-Square Distribution with Ten Degrees of Freedom and the Standard Normal Distribution.

Table 5. Selected Rates of Detection of Aberrant Response Patterns by the Likelihood Ratio Evaluated with True and Estimated Item Parameters

False Pos. Rate	Test	Polychot. MFS True	Polychot. MFS Est.	Dichot. MFS True	Dichot. MFS Est.	3PL Est.
		10 Spuriously Low Responses Per Test				
.001	V	29	26	23	22	19
	Q	18	13	06	06	07
	MT	41	16	23	25	20
.01	V	56	51	38	38	38
	Q	28	27	13	14	15
	MT	69	64	47	47	43
.03	V	68	65	52	52	51
	Q	40	37	23	23	22
	MT	80	76	59	58	59
.05	V	75	73	59	58	57
	Q	48	46	29	28	28
	MT	85	82	66	65	64
.10	V	85	84	71	70	69
	Q	63	60	40	39	37
	MT	91	90	77	76	75
		10 Spuriously High Responses Per Test				
.001	V	02	02	01	01	02
	Q	03	02	02	02	03
	MT	05	05	03	05	04
.01	V	07	07	06	06	06
	Q	12	12	09	10	10
	MT	19	18	13	13	14
.03	V	14	14	12	12	13
	Q	24	22	19	19	18
	MT	32	29	25	24	25
.05	V	19	18	17	15	18
	Q	31	28	26	26	23
	MT	39	36	31	30	32
.10	V	30	28	27	26	27
	Q	43	40	39	37	36
	MT	50	46	43	43	45

however, that the detection rates for both dichotomous analyses fall short of the polychotomous model detection rates. These differences are especially large for the spuriously low response patterns.

Dimensionality misspecification. Table 6 presents results for the misspecification condition in which two-dimensional item responses are analyzed with a one-dimensional model. Results for the correctly specified multi-test analyses are given beneath the columns headed MT.

Substantial drops in rates of detection of both spuriously high and spuriously low response patterns are apparent for all three types of analyses. For example, when the false positive rate is 3% there was a 17% decrease in the rate of detection of spuriously low response patterns by the polychotomous MFS analysis (i.e., 80% detection in the correct analysis versus 63% in the misspecified analysis) and there were 18% decreases for the dichotomous MFS analysis and the three-parameter logistic analysis. A similar pattern of results occurs for the spuriously high response patterns.

The detection rates shown in Table 6 indicate that optimal appropriateness measurement is affected by serious violations of unidimensionality. Specifically, it is clear that detection rates are markedly decreased by combining the simulated verbal and quantitative tests and then performing a unidimensional analysis. This finding underscores the importance of earlier research that developed optimal multi-test appropriateness indices (Drasgow, Levine, & McLaughlin, in press; Levine, in preparation).

Misspecified ability densities. Table 7 presents the results for the response patterns created with ability parameters obtained from truncated chi-square distributions but analyzed with the incorrect assumption that the ability distribution was bivariate normal. A very high degree of robustness to this form of misspecification can be seen in Table 7 for all item response models and both types of aberrant response patterns.

The robustness to ability density misspecifications is a result of the equations for the marginal likelihood of a response pattern given in Equations 3 and 4. From these equations it can be seen that the marginal likelihood is the integral of the product of the conditional likelihood of the response pattern and the ability density. For tests of moderate length or longer, the ability density is ordinarily very flat in relation to the conditional likelihood. For example, the maximum of the normal density is about eight times larger than the minimum density on the interval [-2, 2]. In contrast, the maximum of the conditional likelihood may be 10^{10} or even 10^{40} times larger than its minimum on the same interval (Levine & Drasgow, 1988, p. 170). Consequently, the value of the integral is determined primarily by the conditional likelihood function for tests as long as the verbal and quantitative tests simulated here.

Incorrect specification of the number of aberrant responses. The results for the final form of misspecification are given in Table 8. Here response patterns were generated with either 5 or 15 aberrant responses per test generated; optimal indices were then computed with the correct assumption about the number of aberrant responses or analyzed with the incorrect assumption that 10 item per test were aberrant.

Table 6. Selected Rates of Detection of Aberrant Response Patterns by the Likelihood Ratio with Correct and Incorrect Assumptions about Dimensionality

False Pos. Rate	Polychot. MFS		Dichot. MFS		3PL	
	MT	One Test	MT	One Test	MT	One Test
Data Generated with 10 Spuriously Low Responses Per Test						
.001	41	11	23	05	24	05
.01	69	47	47	20	43	19
.03	80	63	59	41	54	36
.05	85	73	66	51	62	46
.10	91	85	77	67	73	62
Data Generated with 10 Spuriously High Responses Per Test						
.001	05	00	05	02	05	00
.01	19	04	15	05	16	04
.03	32	14	25	13	25	13
.05	39	22	32	20	32	20
.10	50	36	46	32	46	33

Table 7. Selected Rates of Detection of Aberrant Response Patterns by the Likelihood Ratio Evaluated with Correct and Misspecified Ability Densities

False Pos. Rate	Test	Polychot. MFS		Dichot. MFS		3PL	
		Correct	Misspec.	Correct	Misspec.	Correct	Misspec.
		10 Spuriously Low Responses Per Test					
.001	V	31	31	20	19	19	18
	Q	13	11	04	03	04	04
	MT	38	42	20	19	22	20
.01	V	53	54	35	34	35	35
	Q	26	26	10	10	12	11
	MT	66	67	43	41	43	42
.03	V	64	64	49	49	48	47
	Q	40	39	19	19	21	19
	MT	81	80	58	57	57	56
.05	V	73	74	56	57	56	54
	Q	48	47	24	25	27	26
	MT	86	86	67	64	66	64
.10	V	83	83	70	69	68	67
	Q	61	61	38	38	40	39
	MT	94	93	78	78	77	77
		10 Spuriously High Responses Per Test					
.001	V	02	00	02	02	01	01
	Q	02	01	02	01	01	01
	MT	05	00	02	01	03	03
.01	V	08	07	07	06	05	05
	Q	11	08	11	08	10	10
	MT	18	14	14	12	13	12
.03	V	15	15	14	14	11	11
	Q	22	21	21	20	21	19
	MT	31	28	28	26	24	26
.05	V	21	21	18	18	17	17
	Q	32	30	28	26	26	26
	MT	39	38	34	33	31	30
.10	V	31	31	28	27	26	25
	Q	44	42	40	40	38	36
	MT	54	52	49	48	46	45

Table 8. Selected Rates of Detection by the Likelihood Ratio with Correct and Incorrect Specifications of the Number of Aberrant Responses

False Pos. Rate	Test	Polychot. MFS Aberr. Assumption 5	10	15	Dichot. MFS Aberr. Assumption 5	10	15	3PL Aberr. Assumption 5	10	15
Data Generated with 5 Spuriously Low Responses Per Test										
.001	V	19	17		08	07		07	06	
	Q	09	04		01	00		02	01	
	MT	18	15		11	05		09	05	
.01	V	32	27		21	16		19	15	
	Q	15	12		07	06		08	06	
	MT	41	31		26	17		25	17	
.03	V	44	37		32	26		29	23	
	Q	23	21		12	10		13	11	
	MT	53	44		37	28		33	26	
.05	V	51	43		39	33		35	29	
	Q	29	27		16	15		17	14	
	MT	59	51		42	35		40	32	
.10	V	62	57		49	42		48	42	
	Q	39	37		25	22		26	23	
	MT	68	64		54	48		52	45	
Data Generated with 15 Spuriously Low Responses Per Test										
.001	V		45	45		22	26		23	28
	Q		19	23		06	07		09	09
	MT		53	57		25	31		31	34
.01	V		65	69		48	52		43	48
	Q		41	42		18	21		20	19
	MT		83	87		60	65		53	56
.03	V		81	83		64	66		58	61
	Q		55	58		32	32		28	30
	MT		91	92		72	75		67	69
.05	V		86	87		70	73		67	69
	Q		65	67		39	39		37	38
	MT		94	94		79	80		75	78
.10	V		93	93		82	84		79	80
	Q		76	77		51	54		48	50
	MT		97	98		89	91		86	87

Table 8 (concluded)

False Pos. Rate	Test	Polychot. MFS Aberr. Assumption 5	10	15	Dichot. MFS Aberr. Assumption 5	10	15	3PL Aberr. Assumption 5	10	15
		Data Generated with 5 Spuriously High Responses Per Test								
.001	V	00	00		01	01		00	00	
	Q	00	00		00	01		00	00	
	MT	00	00		01	01		00	00	
.01	V	03	03		03	03		03	03	
	Q	05	04		04	04		04	04	
	MT	07	06		07	06		05	05	
.03	V	08	08		08	08		06	06	
	Q	11	10		10	08		10	08	
	MT	15	13		13	11		12	11	
.05	V	12	12		11	11		10	10	
	Q	15	14		13	13		14	13	
	MT	19	18		16	16		17	15	
.10	V	22	20		19	19		17	17	
	Q	24	22		24	22		22	21	
	MT	30	27		28	26		27	24	
		Data Generated with 15 Spuriously High Responses Per Test								
.001	V		05	06		03	04		03	04
	Q		09	12		03	11		07	07
	MT		11	07		06	14		09	10
.01	V		13	14		09	09		12	13
	Q		24	26		16	19		16	20
	MT		35	37		23	27		26	29
.03	V		22	22		17	16		21	21
	Q		39	40		28	31		27	31
	MT		48	51		38	41		38	41
.05	V		28	30		22	23		26	27
	Q		45	48		37	38		35	39
	MT		55	59		46	48		44	48
.10	V		40	42		35	35		36	36
	Q		56	60		51	52		49	52
	MT		68	72		59	60		58	62

Surprisingly modest drops in detection rates were obtained for this form of misspecification. An examination of Table 8 indicates that the least robustness occurred for the response patterns generated with five spuriously low responses per test. At a 5% false positive rate, the drops in detection rates were just 8% for the polychotomous MFS model, 7% for the dichotomous MFS model, and 8% for the 3PL model.

Although further analyses would be needed to corroborate this observation, it appears from Table 8 that a greater degree of robustness is obtained when a response pattern is analyzed with a misspecified number of aberrant responses that is smaller than the actual number of aberrant responses. The converse analysis, in which the misspecified number of aberrant responses is larger than the actual number of aberrant responses, yielded somewhat larger drops in detection rates.

IV. CONCLUSIONS AND DISCUSSION

The major purpose of the research described in this paper was to explore the possibility of using optimal appropriateness indices to address practical testing problems. To this end, it was shown that existing algorithms for evaluating optimal indices could be tailored for a specific problem (i.e., testing the hypothesis that a response pattern with a total test score in a narrow range was obtained honestly or dishonestly) and evaluated the performance of the resulting optimal test. An interrelated set of simulations was also conducted to examine the robustness of optimal tests to violations of assumptions.

There can be little doubt that some examinees may be tempted to cheat when valued outcomes are contingent upon obtaining a test score exceeding some cutoff value. Moreover, the use of cutoffs to determine allocation of valued outcomes is very common: recruitment bonuses, minimum qualification for military enlistment, professional licensing (e.g., nursing, attorney's bar examinations), certification, and state and local public sector hiring.

A way that test administrators can combat cheating has been described in this paper. The statistic given in Equation 8 provides a most powerful test of the hypothesis that an examinee obtained a score barely exceeding some cutoff by honest means against the alternative hypothesis that the barely passing score was obtained by cheating on k items. Of course, the optimal appropriateness index cannot replace careful proctoring during exam administration, routine replacement of old test forms with new test forms, and other security measures. Nonetheless, it does give the test administrator an additional method for identifying cheating. Moreover, test takers may be dissuaded from attempting to cheat if they know that their responses will be examined for indications of cheating.

Tables 1 through 4 give rates of detection of simulated cheaters who obtained scores in a moderately low (20th through 24th percentiles) or just above average (50th through 54th) score ranges. The results given in these tables provide news that is both bad and good. The bad news is that it is very difficult to distinguish between normal response patterns with test scores in a narrow score range and patterns from examinees who cheated on a

few items (5 or 10 per test) in order to obtain test scores in the same range. This result is not too surprising because some of the honest examinees obtained test scores in the given score range by chance rather than merit. Specifically, consider a plot of the frequency distribution of θ or true score for people with observed scores between, say, the 50th and 54th percentiles for some unidimensional test. We would observe many people with θs or true scores that fall outside the 50th through 54th percentiles. The point is that restricting observed scores to lie within some percentile range does not guarantee that θs or true scores will fall in the same percentile range. Some lower ability examinees obtained test scores in the score range because they were lucky and some higher ability examinees obtained test scores in the score range because they were unlucky.

Given just a response pattern, the effects of "luck" (i.e., a few extra correct responses) and the effects of cheating on a few items (again, a few extra correct responses) are very difficult to differentiate. Some of the cheaters have θs in or even above the percentile range. Others have θs just below the percentile range and would therefore have close to a 50% chance of obtaining an observed score in the percentile range if they were retested with a different test form. In sum, there is little practical need to identify cheaters with θs that are close to or in the percentile range, although ethical and policy considerations may deem otherwise.

Turning now to the good news from Study One, Tables 1 through 4 show that it is possible to identify simulated cheating on a relatively large number of items. For the lower test score range, reasonably high rates of detection were obtained with simulated cheating on 10 items per test. Fairly good detection rates were also obtained with cheating on 15 items per test for the just-above-average score range. Identifying individuals who cheat on a large number of items is particularly important because these people have θs that are far below noncheaters.

The results obtained in Study Two clearly suggest that optimal indices can be used effectively in applied settings. Only one form of model misspecification substantially decreased detection rates. This type of misspecification would occur if a test administrator were to combine a verbal test and a quantitative test and treat the composite as a long unidimensional test. Such an event, perhaps based on the argument that typical paper-and-pencil tests are "highly g saturated," would seriously undermine attempts to identify aberrant response patterns. Instead, multi-test optimal appropriateness indices (Drasgow, Levine, & McLaughlin, in press) should be computed because they provide far more effective identification of aberrance in the context of a battery of several unidimensional tests.

Three other forms of misspecification were found to have little or no effect on detection rates in Study Two. Perhaps the most important of these three types of misspecification concerns item parameter estimation errors. In a practical setting, there is never access to the "true" item parameters; at best there are only item parameters estimated from data provided by a large and representative sample. Table 5 shows that there was little decrement in detection rates due to estimation errors for either MFS estimation or 3PL estimation. These results corroborate and extend earlier research on MFS estimation via the ForScore computer program (Drasgow, Levine, Williams, McLaughlin, & Candell, in press; Lim et al., 1989;

Williams & Levine, 1984, 1986) and 3PL estimation with the BILOG computer program (Levine et al., under review; Lim & Drasgow, in press; Mislevy, 1986; Mislevy & Stocking, 1989). It was thus concluded that estimated item parameters can be used effectively in place of the true parameters, provided that the estimates were obtained from a large, representative sample.

Table 7 shows that even a rather badly misspecified ability density has little effect on detection rates, at least for tests of the length simulated in Study Two (50 and 54 items) and the one ability density in this study. This result is convenient because it means that test administrators do not need to be concerned with density estimation. Misspecified ability densities may have a significant effect on shorter tests where the ability density exhibits considerable variation relative to the likelihood function. In such cases it may be necessary to estimate the ability density (see, for example, Levine, 1989a; Mislevy, 1984; or Samejima, 1981).

The final form of misspecification concerned the number of aberrant responses. Table 8 presents the surprising result that an analysis assuming 10 spuriously low responses per test for response patterns that actually had 5 or 15 spuriously low responses per test was almost as effective as the truly optimal analysis. A similar finding was obtained for spuriously high responses. These results provide a contrast between longer, paper-and-pencil tests and short computerized adaptive tests (CATs): Candell and Levine (1989) found larger drops in detection rates when the number of aberrant responses was misspecified on a 15 item CAT.

The results from Studies One and Two lead to the following suggestion for the use of appropriateness measurement in an applied setting. First, the test administrator should make a judgment about the minimum number $\underline{k}$ of spuriously high or spuriously low responses that is needed in order to constitute a nontrivial practical problem. An optimal appropriateness index could be computed assuming $\underline{k}$ aberrant responses, perhaps using existing algorithms and software. Finally, response patterns with index scores that exceed a threshold associated with some acceptable false positive rate could be flagged, and the examinees retested.

Implicit in the above suggestion is the need for item parameters estimated from a large and representative sample. The suggestion also builds on the misspecification analyses that found ability density misspecification to be unimportant and found robustness to misspecification of the number of aberrant responses.

Finally, the utilization of appropriateness indices, perhaps in the manner outlined above, would be expected to improve the quality of a testing program. It would allow identification of some response patterns with modest degrees of aberrance and effective detection of patterns with substantial degrees of aberrance and might thereby deter cheating. It would provide individual test takers with some assurance that their aptitudes had been accurately measured. For these reasons it is recommended that testing programs seriously consider implementing appropriateness measurement.

REFERENCES

Candell, G. R., & Levine, M. V. (1989). Appropriateness measurement for computerized adaptive tests (AFHRL-TP-89-15). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Drasgow, F., Levine, M.V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. Applied Psychological Measurement, 11, 59-79.

Drasgow, F., Levine, M.V., & McLaughlin, M. E. (in press). Multi-test extensions of practical and optimal appropriateness indices. Applied Psychological Measurement.

Drasgow, F., Levine, M.V., McLaughlin, M. E., & Earles, J. A. (1987). Appropriateness measurement (AFHRL-TP-87-6, AD-A184185). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. British Journal of Mathematical and Statistical Psychology, 38, 67-86.

Drasgow, F., Levine, M. V., Williams, B., McLaughlin, M.E., & Candell, G. L. (in press). Modeling incorrect responses to multiple-choice items with Multilinear Formula Score theory. Applied Psychological Measurement, 12.

Levine, M. V. (1985). Classifying and representing ability distributions (Measurement Series 85-1). Champaign, IL: University of Illinois, Department of Educational Psychology.

Levine, M. V. (1989a). Classifying and representing ability distributions (Measurement Series 89-1). Champaign, IL: University of Illinois, Department of Educational Psychology.

Levine, M. V. (1989b). Parameterizing patterns (Measurement Series 89-2). Champaign, IL: University of Illinois, Department of Educational Psychology.

Levine, M. V. (in preparation). Properties of likelihoods of response patterns for short and long tests.

Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. Psychometrika, 53, 161-176.

Levine, M. V., Drasgow, F., Williams, B., McCusker, C., & Thomasson, G. L. (under review). Distinguishing between item response theory models.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. Journal of Educational Statistics, 4, 269-289.

Lim, R. G., & Drasgow, F. (in press). An evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. Journal of Applied Psychology.

Lim, R. G., Williams, B., McCusker, C., Mead, A., Thomasson, G. L., Drasgow, F., & Levine, M. V. (1989). A nonparametric polychotomous model and estimation procedure. Paper presented at the 1989 Office of Naval Research Contractors' Meeting on Model-Based Psychological Measurement, Norman, OK.

Mislevy, R. J. (1984). Estimating latent distributions. Psychometrika, 49, 359-382.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. Psychometrika, 51, 177-195.

Mislevy, R. J., & Bock, R. D. (1984). BILOG II user's guide. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. Applied Psychological Measurement, 13, 57-75.

Rudner, L. M. (1983). Individual assessment accuracy. Journal of Educational Measurement, 20, 207-219.

Samejima, R. (1981). Final report: Efficient methods of estimating the operating characteristics of item response categories and challenge to a new model for the multiple-choice item (Technical Report). Knoxville, TN: University of Tennessee, Department of Psychology.

Sato, T. (1975). The construction and interpretation of S-P tables (in Japanese). Tokyo: Meiji Tosha.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. Psychometrika, 49, 95-110.

Williams, B., & Levine, M. V. (1984). Maximum likelihood for qualitative models. Paper presented at the 1984 Office of Naval Research Contractors' Meeting on Model-Based Psychological Measurement, Princeton, NJ.

Williams, B., & Levine, M. V. (1986). The shapes of item response functions. Paper presented at the 1986 Office of Naval Research Contractors' Meeting on Model-Based Psychological Measurement, Gatlinburg, TN.

Williams, B., & Levine, M. V. (in preparation). ForScore: A computer program for nonparametric item response theory.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116.